Computer-Based Training and Repeated Test Performance: Increasing Assessment Fairness Instead of Retest Effects

Michael Hermes*, Julia Maier, Justin Mittelstädt, Frank Albers, Gerrit Huelmann, and Dirk Stelling

German Aerospace Center DLR, Department of Aviation and Space Psychology, Hamburg, Germany

*Correspondence concerning this article should be addressed to Michael Hermes, German Aerospace Center DLR, Aviation and Space Psychology, Sportallee 54a, 22335 Hamburg, Germany. Email: michael.hermes@dlr.de, Phone: +49 40 513096 44, Fax: +49 40 513096 60

Abstract

When subjects are repeatedly tested in cognitive assessments, systematic score gains occur. Such retest effects become even greater when test preparation is provided between assessments. In the context of personnel selection, retest gains are often increased by commercial test training, which threatens the fairness of psychological testing because not all candidates can afford such offers. In the present study, computer-based training was freely offered to all candidates as part of the personnel selection procedure. We examined the relationship between repeated cognitive ability measurements in high-stakes settings and the amount of computer-based training before each measurement. Analyses of 212 candidates showed that cognitive ability scores and the amount of prior training were only related on the first assessment but not on the second. There were still retest effects, but the magnitude of score gains was negatively correlated with the amount of initial training and was unrelated to training between assessments. Only the change in training amount was positively correlated with retest effects. We conclude that providing all candidates with preparatory training already before the first assessment substantially increases assessment fairness in the personnel selection process.

Keywords: personnel selection, assessment fairness, test preparation, cognitive ability, retest effects, computer-based training

Computer-Based Training and Repeated Test Performance: Increasing Assessment Fairness Instead of Retest Effects

In the context of personnel selection (Lievens et al., 2021; Potočnik et al., 2021), digitization has led to significant changes for organizations and institutions as well as for applicants. In addition to many innovations in selection methods (Woods et al., 2020), the way applicants can prepare for a selection process has changed significantly. This also applies to preparation for cognitive ability tests, which are regularly used in diagnostics in work areas with higher cognitive requirements. There is an abundance of offers for "free online IQ tests", but also specific online forums where test strategies or even test items are shared (Chen et al., 2008; Matton et al., 2009). In many areas, a commercial preparation market has evolved, where software and training tools are made available, that resemble the original ability tests as closely as possible. Especially for jobs with low admission rates, commercial training and coaching providers are very common. Although the conditions during test practice and coaching differ from an actual assessment in a professional setting (subjects can practice more or less seriously, at different times of day, or with various degrees of attention), there is evidence that such practice activities may result in substantial score gains (Hausknecht et al., 2007). A major problem with all these training options is that they are usually not equally available to all applicants, whether due to ignorance, lack of access, or lack of financial resources (Buchmann et al., 2010). Due to this circumstance, only some applicants can benefit from training effects and in turn have an advantage over those who cannot.

A similar problem arises from individual differences in retest experiences. Many organizations allow further participation in selection assessments after an initial negative outcome (Villado et al., 2016). The increasing prevalence of online testing also makes it easier to conduct an assessment for practice purposes only at another organization that uses a similar testing program (information about this is available in

online forums). If organizations are chosen who use online versions of ability tests for initial screening and a second, proctored test administration for final decision-making (Randall & Villado, 2017), the effect is even greater due to the repeated test execution. Since only candidates who had the opportunity to repeat an examination can benefit from the resulting retest effects (Van Iddekinge & Arnold, 2017), they may have an advantage over individuals who conduct an ability test for the first time.

Importantly, organizations and institutions cannot control or reliably measure individual training or retest effects (Scharfen et al., 2018). On the one hand, this makes the psychological selection process considerably more unfair and, on the other hand, it can also threaten the validity of the selection decision (Villado et al., 2016). Several authors have therefore suggested that all applicants should be provided with the same preparation and training opportunities (Arendasy et al., 2016; Freund & Holling, 2011; Levacher et al., 2021; Sackett et al., 1989; Villado et al., 2016). Villado et al. (2016), for example, recommended "to provide test-takers with a degree of test practice so that all or most test-takers achieve maximum test-specific knowledge prior to completing the assessment that will be used for decision making" (p. 245). However, such preparation possibilities are still rarely provided (Hermes et al., 2019). Accordingly, there is still hardly any literature in this area. Exceptions are Hermes et al. (2019), who investigated measurement invariance in computer-based training, and Campion et al. (2019) who offered retired versions of knowledge tests for training purposes prior to their assessment (which could be carried out online up to two times). However, it remains an open question how free test-specific training is associated with repeated cognitive ability measurements and retest effects—especially in high-stakes settings such as personnel selection.

In the present study, all candidates received computer-based training as part of the (high-stakes) personnel selection process, which allowed repeated testing. This enabled us to investigate how the amount of digital, test-specific training opportunities is related to performance in repeated assessments and whether the associations between training and performance differ from initial to subsequent test administrations—which is still unclear.

Computer-Based Training and the Initial Cognitive Ability Measurement

When candidates prepare a *first* high-stakes assessment with test-specific training, it can be expected that the amount of training is positively correlated with the test scores of the cognitive ability assessment. This expectation is based on evidence from studies on working memory training (e.g., Jaeggi et al., 2014), coaching effects (e.g., Messick & Jungeblut, 1981), and studies that did not employ training tools but employed up to ten testing sessions (Albers & Höft, 2009; Bartels et al., 2010). These studies consistently found an increase in cognitive performance across training or testing sessions, respectively. In addition, using test-specific training modules in a high-stakes selection context, Hermes et al. (2019) reported a mean correlation of $r = .39$ between the number of training runs and test performance in a sample of 15,752 applicants.

The factors underlying this consistent positive relationship can be derived from models of retest effects (Lievens et al., 2007; Randall & Villado, 2017; Van Iddekinge & Arnold, 2017). According to this line of research, score gains in cognitive ability tests are mainly attributable to two clusters of effects: a decrease in interfering factors (discomfort, confusion, unfamiliarity with the test) and an increase of test-specific skills. In addition to a number of moderator variables, there is also the possibility of true ability increases, but these are only relevant in the knowledge domain. Recent evidence

suggests that the second cluster (increase in test-specific skills) plays a central role in the generation of retest effects (e.g., Arendasy & Sommer, 2017; Hayes et al., 2015). These studies have demonstrated the importance of learning processes during the execution of cognitive ability tests. In particular, when subjects repeatedly work through a cognitive ability test, information on speed and accuracy of the solution strategy and salient item design characteristics are memorized (Arendasy & Sommer, 2017). With further retests, the solution strategy becomes more automated and hence more working memory resources become available to improve the overall strategy. For the use of freely available computer-based training already employed before the initial assessment, it can be expected that the development of solution strategies already takes place during the training sessions and to a lesser degree within the actual assessments. With regard to the positive relationship between training/testing sessions and test performance, as well as general models of cognitive learning processes (Donner & Hardy, 2015), the use of computer-based training will result in a learning curve, with the learning effect being largest in the first sessions and becoming smaller with further training. On average, candidates who train extensively should reach their personal learning plateau, while those who train less might miss some potential for training gains. A similar pattern can be partially assumed for the first cluster of causes of score gains: confusion and unfamiliarity with the test will most strongly decrease during the first training runs and are only of minor relevance from the first assessment to the second.

Hypothesis 1: The amount of computer-based training before a *first* high-stakes assessment is positively correlated with cognitive ability scores from this first assessment.

Computer-Based Training and the Second Cognitive Ability Measurement

When candidates have the opportunity to be retested and prepare again with test-specific training before the assessment, the situation becomes more complex. Candidates starting the second training phase now differ not only in their cognitive ability but also in their experience from the first assessment and their actual level of test preparation. The level of test preparation before the second training phase depends on a variety of factors such as the amount of training before the first assessment, the time interval since the first assessment, or their general ability to recall solution strategies. In addition, there will be motivational differences between candidates depending on their pass/fail result of the first assessment (Barron et al., 2017). In the case that candidates have no feedback about which tests were crucial for a negative assessment result, there will also be differences between candidates in the correctness of their hypotheses about which ability tests may have been crucial and should therefore be prepared for the most. Importantly, with regard to the learning processes outlined above (Bartels et al., 2010; Donner & Hardy, 2015), candidates approach their personal learning plateau during the first training phase. Since these training effects can be expected to last over longer periods of time (Uttal et al., 2013), a second training phase has less potential for learning effects than the first one. As a result, performance in the second assessment should be related to the amount of previous training to a lesser extent than performance in the first assessment.

Hypothesis 2: The relationship between cognitive ability scores and the preceding amount of computer-based training is smaller at the second high-stakes assessment than at the first high-stakes assessment.

Computer-Based Training and Retest Effects

There is a great deal of evidence that retest effects in cognitive ability tests occur if no training or practice sessions were carried out at all (Hausknecht et al., 2007; Scharfen et al., 2018). In addition, there are several studies that assessed the impact of training and coaching activities carried out between the first and second test administration but with no training before the initial one (e.g., Freund & Holling, 2011; Hausknecht et al., 2007; Levacher et al., 2021). However, we found no research in the selection context where (free) test training was used prior to all assessments. Studies with training only between test administrations have typically found that retest gains were greater when training was conducted compared to when no training was conducted. Freund and Holling (2011), for example, repeatedly used figural matrices tests in a group of students, with about half of the participants having daily online training between test administrations and the other half having no training opportunities. They found that the group with training between sessions had a retest effect that was almost twice as large as in the group without training (with training: $d = 0.97$, without training: $d = 0.49$, averaged across identical and parallel test versions). For the use of freely available computer-based training already employed before the initial assessment, it can be expected that retest effects become smaller because some of the underlying factors of retest effects (e.g., development of solution strategies, reduction of misunderstandings) become effective during the initial training phase (see above). Villado et al. (2016) noted that "the provision of sufficient practice (…) may limit retest effects attributable to test-specific knowledge and skills" (p. 245).

For the following reasons it can be expected that retest effects do not disappear completely after initial training: the proportion of candidates who reach their individual learning plateau can be expected to be greater after the second than after the first

training phase. In addition, factors such as discomfort with the situation and stress reactions are more likely to be reduced by repeated participation in the real assessment setting, which is particularly different from training settings when personnel selection contexts are considered. These factors may interfere less in the second assessment if some habituation to the test setting has occurred (Grissom & Bhatnagar, 2009). As a result, ability scores are greater in these candidates in their second compared to their first assessment.

Currently, there is no study that explicitly examines retest effects following the use of free test training prior to all assessments. There is some evidence for existing retest effects after the use of computer-based training before the initial assessment in a study of Hermes and Stelling (2016). The authors analyzed the impact of occasion-specific effects on ability measurements in the framework of latent state-trait theory and reported retest effects as an additional finding ($d = 0.47$, calculation based on their Table 1). However, the amount of training was not reported and it is unclear to what extent the training modules were used prior to the first and second assessment.

Hypothesis 3: Retest effects occur even after initial training.

If retest effects are still present, the question arises how strong these score gains are related to the amount of test training. As described above, when training is only conducted between test sessions, then retest effects increase if training is used compared to settings where training is not used (e.g., Freund & Holling, 2011; Levacher et al., 2021). That is, training and retest effects are correlated. However, if free training opportunities are offered even before the initial assessment, confusion and unfamiliarity with the test will most strongly decrease during the first training phase. In addition, solution strategies can be developed during this first training phase so that this factor can no longer contribute to the development of retest effects to such a significant extent:

the more candidates train and the more solution strategies are developed and automated, the less potential for score improvements remains. A similar rationale was developed by Scharfen et al. (2018) to account for decreasing retest effects in studies with more than two test administrations. They hypothesized that a reduction of distorting factors and an improvement of test-specific skills will most likely take place "within the first sessions and not afterwards" (pp. 56f). Since these processes can be expected to take place before the first assessment, if training was conducted prior to the first assessment, and to last over time (Uttal et al., 2013), we hypothesized that a second training phase between the assessments and the amount of this training is not related to the size of the retest effects.

Hypothesis 4a: Retest effects are negatively related to the amount of computer-based training conducted before the *first* assessment.

Hypothesis 4b: Retest effects are not related to the amount of computer-based training conducted *between* assessments if training was also conducted before the first assessment.

All hypotheses imply that the training is highly test-specific, that is, item characteristics and test principles are as similar as possible in training modules and actual ability tests.

Method

Participants and Procedure

The sample consisted of $N = 212$ licensed airline pilots applying for a pilot position at different major European Airlines. Candidates were included in the sample if they had participated in at least two assessments as a licensed pilot applicant in our institution. The candidates were between 19 and 51 years old ($M = 28.17$, $SD = 6.00$)

when they participated in their initial assessment; 88% were male, 12% were female. All of them had completed high school education adequate for university entrance. Individuals who had undergone a prior assessment in our institution, e.g. as a pilot trainee or an air traffic controller trainee applicant, were excluded. Data sets with missing data were also excluded. All data were collected between 2010 and 2019.

The data presented in this study were derived from the first step of a multi-stage personnel selection procedure. Details of the whole selection procedure and possible subsequent stages of assessment for those candidates who passed the first step can be found in Zinn et al. (2020). The assessments were conducted in a highly standardized environment. All tests were administered in a large, well-lit and air-conditioned testing room with up to 44 candidates. Every test started with computerized instructions, followed by a few sample and practice items. Before the main test began, questions concerning comprehension of the test procedure were answered by the test administrator. A few days after the assessment, candidates were informed about their assessment outcome (pass/fail). However, they were not informed about the results of single ability tests and, in the case of a negative outcome, which tests were decisive for this outcome. Of all candidates, 40 (19%) passed the initial assessment, 172 (81%) failed it. It must be noted, however, that a negative outcome could also be a result of tests taken during the assessment but not analyzed in this study. In only about one third of all failed candidates, the negative outcome of the initial assessment resulted from at least one of the five ability tests analyzed in the present study. Therefore, a negative assessment outcome must not be equated with a low performance in these five ability tests. In the second assessment, candidates had to perform all ability tests again, irrespective of their outcome of the first assessment. The time span between the first and the second assessment was between 6 and 2,604 days (M = 394, Mdn = 197, SD = 516).

The pass rate of the present sample (N = 212; 19%) was lower than the pass rate of all candidates who had applied as licensed airline pilots and had participated in an initial assessment between 2010 and 2019 in our institution (regardless of if they had participated in a retest; N = 2012; pass rate: 45%). This is not surprising, as not all candidates with a positive assessment decide to participate in an assessment again. There were several reasons why a candidate may have chosen to retest despite a positive initial result: a negative result in the later stages of the selection procedure; the intention to change the employer (airline), or a parallel application to a different employer.

Measures

For the present study, we analyzed five computer-based tests, designed to measure the basic cognitive abilities of visual perception speed, selective attention, auditory and visual memory capacity, and mental rotation. These tests were developed in our institution and have been used over a long period of time during the first stage of pilot selection. In the selection process, they were part of a larger test battery of 12 tests in total, which also included tests for psychomotor control and multiple task capacity, as well as occupation-specific multiple-choice knowledge questionnaires, personality and attitude inventories. The five ability tests were selected for this study because they were administered in identical versions to all candidates for all measurements. The sequence of presentation was also identical for these tests. In addition, the five tests were subject to the same test preparation concept. The remaining tests of the test battery did not meet at least one of these selection criteria and were therefore not considered. The following cognitive ability tests were examined (indicators of test-retest reliability are presented in Table 1):

The *Optical Perception Test* is a measure of visual perceptual speed that requires rapid visual search and the correct reading of four specific instrument dials

within a complex display of nine instruments, which is presented only for a very short time. The instruments differ in features such as color and shape. Prior to every task presentation, brief information is displayed that specifies one of the features as being critical. Thereafter, all four instruments matching this specific critical feature must be found and read correctly. A fixed number of tasks must be completed and the test duration is about 12 minutes.

The *Symbol Concentration Test* is a measure of selective attention that requires the application of rules that change at specific intervals to long sequences of displayed triangles. The triangles differ in certain visual features and before each sequence is displayed, a rule specifies which of these features are to be considered critical. For the entire sequence it must then be decided in each case whether or not two consecutive triangles are identical with respect to these critical features. The self-paced answering procedure is limited to 13 minutes of testing time.

The *Running Memory Span Test* is a measure of auditory memory that requires memorization of acoustically presented numerical sequences of varying length that end unpredictably. Subsequently the memorized digits must be reproduced in reverse order starting with the last digit presented. The test contains a fixed number of items and takes about 20 minutes.

The *Visual Memory Capacity Test* consists of an n-back task requiring the memorization and comparison of abstract symbols that differ in shape and color and are presented visually in a long sequence one after the other. It must be decided for each stimulus, whether the displayed symbol is identical or different to the one that was presented n steps earlier in the sequence. In the course of the test, 6 time-limited runs of 2-back, 3-back, 4-back, and 5-back sequences are administered that take 20 minutes in total.

In the *Mental Rotation Test*, a cube with one face marked by a cross must be visualized and thereafter—following a sequence of acoustic commands—mentally rotated. The sequence of commands varies in length and presentation speed. At the end of each task the final orientation of the cube must be determined. The test contains a fixed number of items and takes about 25 minutes.

Training and Test Preparation

To ensure that all candidates had the opportunity to prepare adequately for the assessments, everyone received free online access to computer-based training modules for at least 20 days before their scheduled testing day. These training modules reflected the respective ability tests as closely as possible and had to be downloaded for offline use. Every training module contained different parts: first, the specific test principle was explained in an instructional text. Afterwards, practice sessions were offered at different levels of difficulty. Users were advised to start their training at the lowest level of difficulty and gradually increase it as their performance improved. At the highest level, the task difficulty was equivalent to the actual test, yet without disclosing any original test items. For variation of contents between repeated executions of the training programs, items were either randomly drawn during the runtime from a larger pool of items or dynamically generated, thereby following specific rules to control item difficulty. At the end of each training run, candidates received summary feedback on the percentage of items answered correctly to evaluate their own performance. Candidates were also provided with general training instructions. It was recommended that users should not repeat the same module more than three times in a row but rather keep the training varied by using training modules for different abilities alternately. Regular breaks should also be included in the training schedule. The recommendation was to complete each module 20 times. If thereafter candidates had the impression that

their performance was still increasing, they were advised to practice even more. All candidates were provided with a log sheet to be completed during the training phase. This document was not only useful for individual training monitoring, but also part of their official application documents, which had to be submitted in full on the examination day. However, only the number of completed training runs had to be recorded, not the individual performance scores. According to our training concept, it is important that all candidates can carry out the training "unobserved" and free of any pressure, which could be caused by a falsely assumed performance control. Therefore, training performance scores were not recorded and accordingly were not available for further analysis. For the present study, the number of training runs for each ability test was taken from these log sheets and averaged over the five relevant ability tests to test our hypotheses.

Statistical Analyses

We converted the raw scores from each ability test into *T*-scores, i.e. standardized scores with a mean of 50 and a standard deviation of 10. The mean raw score aggregated across both assessments and the pooled standard deviation were used for the *T*-transformation. In the next step, Cohen's *d* was computed for each ability test separately with the formula: $d = (M_2 - M_1)/SD_{pooled}$, where M_i is the mean *T*-score of the first (T1) or second (T2) assessment, respectively, and SD_{pooled} is the pooled standard deviation. For the following analyses, we used *T*-scores averaged across the five ability tests and the mean number of training runs averaged across the five corresponding computer-based training modules. Retest effects were operationalized as difference scores of the two measurements (T2 minus T1). We preferred the difference method to the residual method because the use of the residual method entails the risk of bias (Castro-Schilo & Grimm, 2018).

When retest effects were predicted, the training amount before the first and second assessment, the outcome of the first assessment (pass/fail), and the retest interval (in days) were included as predictors in a multiple linear regression analysis. The retest interval (Hausknecht et al., 2007; Scharfen et al., 2018) and the outcome of the first assessment (Randall & Villado, 2017) were included as controlling factor because of its possible influence on the size of retest effects. We carefully checked the assumptions underlying linear regression analysis (linearity of the phenomenon measured, independence and constant variance of the error terms, normality of the error term distribution, and multicollinearity) according to Hair et al. (2018).

Statistical analyses were performed with IBM SPSS Statistics (Version 26). For comparisons between correlation coefficients, we used the R based package cocor (Diedenhofen & Musch, 2015).

Results

In Table 1, descriptive statistics for the amount of training and the resulting test scores from all tests are presented for the first and the second assessment. In all tests, the performance scores were higher in the second than in the first assessment. Retest effects (i.e., changes in performance scores) ranged between Cohen's $d = 0.30$ and 0.51 for individual ability tests, with a mean of $d = 0.37$. In contrast to the performance results, the number of training runs decreased from the first to the second assessment in all measures. Of all candidates, 95% had completed each module on average at least eight times before the first assessment and at least seven times before the second. There were no applicants who had not used the training at all.

Table 1

Descriptive Statistics for the Number of Training Runs and Test Performance in the First and Second Assessment

	Number of training runs		Test performance (*T*-scores)		Retest effect	Test-retest reliability
Assessment:	1	2	1	2	2-1	1↔2
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*d*	*r*
Perceptual speed	25.69 (14.16)	22.49 (11.82)	48.13 (10.58)	51.87 (9.38)	.37	.71
Selective attention	18.73 (9.25)	17.26 (9.73)	47.46 (11.34)	52.54 (8.45)	.51	.81
Auditory memory	22.33 (11.13)	20.13 (10.72)	48.44 (9.96)	51.56 (10.04)	.31	.74
Visual memory	20.99 (10.74)	19.34 (11.16)	48.13 (10.95)	51.87 (8.95)	.37	.85
Mental rotation	20.55 (10.83)	19.08 (11.13)	48.49 (11.83)	51.51 (7.76)	.30	.66

Descriptive statistics and intercorrelations of performance and training variables averaged across the five ability domains are presented in Table 2. The amount of computer-based training before the first assessment (training before T1) was significantly correlated with the cognitive ability scores of the first measurement (performance T1, $r = .27$, $p < .001$). The more candidates had practiced with the training modules, the higher were their achieved test results. *Hypothesis 1* was thus supported. In contrast, the amount of training immediately before the second assessment (training before T2) was not correlated with cognitive ability scores of the second measurement (performance T2, $r = .01$, $p = .875$). The difference between the two correlation

coefficients (Δr = .26) was statistically significant, the confidence limits did not include zero (Zou, 2007), 95% CI [0.11, 0.41]. *Hypothesis* 2 therefore was also supported.

Table 2

Descriptive Statistics and Intercorrelations of Performance, Training, and Outcome Variables

	M	*SD*	Skew-ness	Kurtosis	1	2	3	4	5	6	7
1. Performance T1	48.13	8.54	-1.36	2.64	–						
2. Performance T2	51.87	6.57	-1.27	2.73	.84***	–					
3. Performance change	3.74	4.62	1.05	2.12	-.65***	-.14*	–				
4. Training amount T1	21.66	10.17	1.15	1.44	.27***	.21**	-.20**	–			
5. Training amount T2	19.66	9.78	1.59	3.46	-.04	.01	.10	.42***	–		
6. Training change	-2.00	10.79	-0.03	2.80	-.30***	-.19**	.27***	-.57***	.52***	–	
7. Retest interval (days)	393.74	516.33	2.26	5.47	-.14*	-.16*	.04	.05	.17*	.10	–
8. Pass/fail T1	–	–	–	–	.39***	.32***	-.26***	-.10	-.17*	-.06	-.04

Note. All correlations are based on performance and training variables averaged across the 5 ability domains. Change in performance (i.e., retest effects) and training were calculated by subtracting scores of the first assessment (T1) from the second (T2). A fail was coded as 1 (81%), a pass as 2 (19%).

*p < .05. **p < .01. ***p < .001.

Hypothesis 3 stated that retest effects occur even after initial training. The descriptively found retest effect ($d = 0.37$) was further analyzed with a paired sample *t*-test. In line with *Hypothesis* 3, the difference in *T*-scores between the first and second assessment was statistically significant, $t(211) = -11.79$, $p < .001$. In an exploratory analysis, we investigated whether this retest difference could be explained by regression to the mean (Bobko, 2001, p.167). Since the pass rate in the current sample was smaller than in the total of all applicants (19% vs. 45%), regression to the mean is a possible cause of the retest difference. The strong correlation between performance T1 and the amount of performance change, $r = -.65$, could also be interpreted as an indication of regression to the mean. When the current sample ($N = 212$) and the total of all applicants ($N = 2012$; see Participants and Procedure) were used in this analysis, the effects of regression to the mean were close to zero for all ability tests (mean $\Delta d = .00$, max. $\Delta d = |.02|$) and thus not influential for the occurrence of retest effects in our sample. In addition, it should be noted that the relationship between performance T1 and the amount of performance change ($r = -.65$) is likely be overestimated due to a potential bias when initial values and change values are correlated (Tu & Gilthorpe, 2007). After a correction of this bias with Oldham's method (Tu & Gilthorpe, 2007), the correlation dropped to $r = -.44$.

In a next step, the relationship between the amount of computer-based training, outcome T1, and retest effects (performance change) was analyzed (*Hypotheses 4a and 4b*). As mentioned before, all candidates had used the training modules before the first assessment. Table 2 shows that performance change was negatively correlated with the training amount before the first assessment ($r = -.20$, $p = .004$) and the outcome (pass/fail) of the first assessment ($r = -.26$, $p < .001$). However, there was no significant

correlation between performance change and training before the second assessment (r = .10, p = .162). To analyze the specific contribution of each variable within a comprehensive model, a multiple linear regression analysis was calculated, including the training variables, outcome T1, and retest interval for the prediction of retest effects. Regarding the assumptions underlying multiple linear regression analyses, residual plots revealed that linearity of the relationship was given. Error terms had constant variances and showed no autocorrelation (Durbin-Watson = 1.87). There was no multicollinearity in the data set: the highest intercorrelation of predictors was r = .42, tolerance (M = 0.89, SD = 0.10) and VIF statistics (M = 1.13, SD = 0.12) were well within critical limits (Hair et al., 2018). Since error terms were not normally distributed, bootstrapping methods (with 1000 samples) were used to obtain robust standard errors. As Table 3 shows, training before T1 and training before T2 both reached statistical significance within the regression equation: The higher the amount of training before T1, the smaller was the performance change. In contrast, the higher the amount of training before T2, the greater was the performance change. It should be noted though, that since the predictors in the multiple regression analysis (such as training before T1 and T2) are partial regression coefficients, training before T2 must be interpreted as residual training change, which is consistent with the significant correlation between (simple) training change and retest effects presented in Table 2 (r = .27, p < .001). *Hypothesis 4a* and *4b* were supported. The outcome T1 (pass/fail T1) also had a significant negative weight indicating that retest effects were larger for candidates who had failed their first assessment. The retest interval had no meaningful regression weight for the prediction of performance change.

Table 3

Prediction of Performance Change by Training Amount, Retest Interval, and Assessment Outcome

	OLS estimates						Bootstrap estimates	
	B	*SE B*	ß	*t*	*p*	R^2	*SE B*	*p*
Training before T1	-0.14	0.03	-.30	-4.19	<.001		0.03	.001
Training before T2	0.08	0.03	.17	2.41	.017		0.03	.013
Retest interval	0.00	0.00	.01	0.17	.863		0.00	.866
Pass/fail T1	-3.10	0.77	-.26	-4.03	<.001		0.56	.001
						.15		

Note. T1 = first assessment; T2 = second assessment; fail was coded as 1, pass as 2.

Discussion

The advancing digitization in test development makes it likely that digital preparation options will be offered and used. In high-stakes selection contexts, assessment fairness and the validity of selection decisions are threatened if some candidates can benefit from (commercial) test preparation and/or retest effects, but others cannot (Hermes et al., 2019; Villado et al., 2016). Although many authors recommended offering free test preparation to all candidates, to our knowledge this is the first study where such training opportunities were realized in repeated assessments in a high-stakes selection setting. We found that the amount of computer-based training before a first assessment was positively correlated with cognitive ability scores of the first assessment and that the relationship between training and test results was significantly smaller at the second compared to the first assessment. There were still retest effects, but the size was negatively correlated with the initial training amount, unrelated to the second training amount and positively correlated with training change. In addition, we found greater retest effects in candidates failing compared to candidates passing the first assessment.

Computer-Based Training and Cognitive Ability Measurements

Previous research has shown that the amount of test preparation is positively correlated with the scores of cognitive ability tests (Hermes et al., 2019). The present study replicates this finding and also shows that this relation is no longer present after the first assessment. The decrease in the correlation of training and test performance from the first to the second assessment is consistent with the development of solution strategies and an increasing test familiarization during the first training phase and a preservation of the learning gains to the second assessment (Uttal et al., 2013).

It could be hypothesized that high-ability individuals do more training runs or benefited from training more than low-ability individuals, also known as rich-get-richer effect (cf. Randall & Villado, 2017). However, recent meta-analytical evidence on cognitive training effects suggests that individual differences in training gains follow a poor-get-richer effect rather than a rich-get-richer effect (Traut et al., 2021): participants with lower cognitive ability gained most from trainings of cognitive ability, implying a compensation rather than a magnification effect of training. Although in the present study the negative correlation between performance in the first assessment and performance change can be interpreted as evidence of such a poor-get-richer effect (r = -.65, bias-corrected: r = -.44), this cannot be clearly concluded because we had no pretraining measurements of cognitive ability.

Computer-Based Training and Retest Effects

Previous research has shown that retest effects increase if test preparation is used between two test administrations (e.g., Freund & Holling, 2011; Levacher et al., 2021). This is important for high-stakes settings because candidates with a negative first assessment may be especially motivated to spend money on commercial test preparation before the second assessment. The present study additionally included test training prior

to a first assessment and showed that in this case, the size of retest effects is negatively correlated with the amount of initial training. This suggests that the more candidates had trained, the less familiarity with the tests and the development of solution strategies could play a role in generating retest effects. In contrast, the second training phase was not associated with retest effects, but a change in training amount was: the more the training was increased or the less it was decreased compared to the first training phase, the higher were the retest effects. This suggests that only those candidates who invested relatively more time in the second training were able to benefit from training effects.

Since it was not possible to implement a control group without initial training, it could also be possible that the non-existent association of the second training with test performance is not due to the initial training, but solely to the fact that all candidates were familiar with the tests due to the initial assessment. According to this hypothesis, the results would also be expected without initial training. However, the following evidence suggests that this is at least very unlikely: First, retest studies with training only between test administrations consistently found an effect of training even though candidates were already familiar with the tests based on the initial testing (e.g., Freund & Holling, 2011; Levacher et al., 2021). If familiarity with the tests based on initial testing was critical, the training should not have resulted in a greater retest effect. Second, in retest studies with up to ten test sessions, an increase in test performance was found even after the first retest (Albers & Höft, 2009; Bartels et al., 2010). This contradicts the notion that test familiarity from a single test administration could be crucial because in this case, it would be expected that there is no further increase in test performance after the first retest. Third, in the present study the retest effect was negatively correlated with the amount of training prior to the first testing, which would not be expected if familiarity with the tests based on initial testing was critical. Taken

together, this evidence suggests that the present results are more consistent with the conclusion that the initial training, rather than test familiarity from the initial assessment, is critical to the lack of effect from the second training.

Our study also showed that retest effects were present even after using test-specific computer-based trainings prior to the first assessment. The size of the retest effects (d = .37) was comparable to studies that did not use computer-based training before the first test administration (Hausknecht et al., 2007; Kulik et al., 1984; Scharfen et al., 2018) but smaller than retest effects in comparable high-stakes settings and with comparable ability tests (Matton et al., 2009, 2011). For example, Matton et al. (2009) reported retest effects of d = 0.85 – 1.02 (for ability tests without a focus on knowledge) in a sample of applicants to a flight school. Since changes in the latent ability are improbable with cognitive ability tests (Scharfen et al., 2018) and effects of solution strategies and test familiarization are limited with free training opportunities, other causal factors of retest effects become more crucial, such as the emotional involvement. As discussed in the retest literature (Van Iddekinge & Arnold, 2017), factors such as discomfort with the situation and stress reactions may be reduced in a repeated high-stakes assessment because candidates "have been through the process and know what to expect" (p. 49). In a less strained state, candidates can allocate more attention to the task—especially after a mistake has been noticed—and solution strategies may be retrieved more effectively (Grant et al., 1998), resulting in a better test performance.

The regression analysis showed that retest effects were greater in candidates failing compared to candidates passing the first assessment, i.e., they were greater in redemptive compared to non-redemptive re-testers. The redemptive status is one of the moderating factors of retest effects that has been little studied (Randall & Villado, 2017). There is only one study that reports retest effects as a function of assessment

outcome and the present result is consistent with this research (Hausknecht, 2010). Differences in retest effects between redemptive and non-redemptive re-testers in high-stakes settings may be attributable to several factors: differences in assessment preparation, ceiling effects in passing candidates, or a higher impact of interfering factors at the first assessment among the failing candidates that could then decrease at the second assessment.

We did not observe a significant correlation between the magnitude of the retest effect and the size of the retest interval. It is possible that solution strategies for specific tests are forgotten over long periods of time. The fact that our candidates took the test-specific training again before the second assessment could reactivate previous solution strategies and thus counteracted the influence of memory decay. Therefore, increasing the length of the retest interval may not be an effective method to reduce retest effects when preparatory training is used.

Although the focus of this study was not on individual difference variables, it should be noted that personality traits could also contribute to the size of retest effects (Randall & Villado, 2017). In the study of Barron et al. (2017), for example, high emotional stability predicted retest outcomes beyond initial test scores for cognitive ability retesting.

Practical Implications

The present study confirms the positive relationship between test preparation and cognitive performance but also suggests that at least in a high-stakes context, such a relationship can only be expected with the first test administration. After a first test session, the relevance of training decreases if training has already been provided prior to the first test administration. Thus, when evaluating training effects on cognitive

performance, it must always be considered if there had been prior assessments with these measures or whether this was the first experience with the ability test.

Since (commercial) training effects that are only available to some candidates can lead to a considerable reduction in assessment fairness and even to incorrect selection decisions, we recommend, as other authors have done (Arendasy et al., 2016; Freund & Holling, 2011; Levacher et al., 2021; Sackett et al., 1989; Villado et al., 2016), that in high-stakes contexts all participants be provided with freely offered preparation opportunities that are as similar as possible to the actual ability tests. This should already be taken into account during test development. Especially in the case of digital ability tests, a training version can easily be created in addition to the actual test. Previous studies have already shown that training does not threaten measurement invariance (Hermes et al., 2019) and that repeated testing does not reduce validity (Van Iddekinge & Arnold, 2017).

The issue of assessment unfairness and the risk of false diagnostic decisions may be a problem not only in the context of personnel selection but also in other diagnostic areas. For example, in follow-up diagnostics in clinical settings (Calamia et al., 2012) or in repeated licensing assessments (e.g., to regain the driver's license) knowledge about training effects is necessary to distinguish between different performance trajectories after training. If the influence of training effects is not known, diagnostic errors may result because decomposition processes may be underestimated or increases in performance overestimated.

Limitations and Recommendations for Future Research

The present study did not include a control group without any preparatory training. Such a control group would allow researchers to determine more clearly to what extent retest effects decrease after initial training. However, in high-stakes

settings, it is ethically unacceptable to equip some of the applicants with preparatory training while others go into a cognitive ability assessment with different or no preparation at all. Future studies could implement a control group with no training at least in a low-stakes context—even if this reduces ecological validity compared to a high-stakes design.

To learn more about the effects of cognitive ability on training behavior and the relationship between initial training and test performance, it would be important to consider cognitive ability scores measured before the first training session. It would also be insightful to have performance scores of the training in addition to the mere number of training runs. Future research should attempt to collect the respective data and provide an analysis of actual training performance.

Finally, the present study examined tests that measure different facets of general mental ability, which are used in selection for many operational professions (Salgado, 2017). However, in many other selection contexts, especially in educational settings, knowledge-based ability tests are used (e.g., SAT or GRE). These tests differ from cognitive ability tests in the aspect that in knowledge-based ability tests, there is the possibility that abilities actually change (e.g., through training). Since such training effects may be different compared to selection contexts where cognitive ability tests (measuring facets of general mental ability) are used, the results of the present study cannot be generalized to settings with a focus on knowledge tests. Therefore, the use of free training options should also be investigated in knowledge-based tests in the future.

Conclusion

Our results suggest that providing preparatory training to all candidates already before the first assessment has the potential to reduce retest effects in cognitive ability assessments—which may be the result of individual (commercial) training activities.

This approach increases the fairness of the selection process and helps to prevent false diagnostic decisions, which may occur in a high-stakes personnel selection caused by the selective use of (digitally) available training tools and strategies.

Disclosure of interest

The authors report no conflict of interest.

References

Albers, F., & Höft, S. (2009). Do it again and again. And again? *Diagnostica, 55*(2), 71-83. https://doi.org/10.1026/0012-1924.55.2.71

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence, 62*, 89-98. https://doi.org/10.1016/j.intell.2017.03.003

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55*, 44-56. https://doi.org/10.1016/j.intell.2016.01.004

Barron, L. G., Randall, J. G., Trent, J. D., Johnson, J. F., & Villado, A. J. (2017). Big five traits: Predictors of retesting propensity and score improvement. *International Journal of Selection and Assessment, 25*(2), 138-148. https://doi.org/10.1111/ijsa.12166

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience, 11*, 118. https://doi.org/10.1186/1471-2202-11-118

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management*. SAGE Publications.

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces, 89*(2), 435-461. https://doi.org/10.1353/sof.2010.0105 %J Social Forces

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The*

Clinical Neuropsychologist, 26(4), 543-570. https://doi.org/10.1080/13854046.2012.680913

Campion, M. C., Campion, E. D., & Campion, M. A. (2019). Using practice employment tests to improve recruitment and personnel selection outcomes for organizations and job seekers. *Journal of Applied Psychology, 104*(9), 1089-1102. https://doi.org/10.1037/apl0000401

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships, 35*(1), 32-58. https://doi.org/10.1177/0265407517718387

Chen, S. Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*(Pt 2), 471-492. https://doi.org/10.1348/000711007x227067

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One, 10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review, 22*(5), 1308-1319. https://doi.org/10.3758/s13423-015-0811-x

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence, 39*(4), 233-243. https://doi.org/10.1016/j.intell.2011.02.009

Grant, H. M., Bredahl, L. C., Clay, J., Ferrie, J., Groves, J. E., McDorman, T. A., & Dark, V. J. (1998). Context-dependent memory for meaningful material: Information for students. *Applied Cognitive Psychology, 12*(6), 617-623.

https://doi.org/10.1002/(sici)1099-0720(1998120)12:6<617::Aid-acp542>3.0.Co;2-5

Grissom, N., & Bhatnagar, S. (2009). Habituation to repeated stress: Get used to it. *Neurobiology of Learning and Memory, 92*(2), 215-224. https://doi.org/10.1016/j.nlm.2008.07.001

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2018). *Multivariate data analysis*. Cengage Learning, EMEA.

Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology, 63*(2), 299-324. https://doi.org/10.1111/j.1744-6570.2010.01171.x

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373-385. https://doi.org/10.1037/0021-9010.92.2.373

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1-14. https://doi.org/10.1016/j.intell.2014.10.005

Hermes, M., Albers, F., Böhnke, J. R., Huelmann, G., Maier, J., & Stelling, D. (2019). Measurement and structural invariance of cognitive ability tests after computer-based training. *Computers in Human Behavior, 93*, 370-378. https://doi.org/10.1016/j.chb.2018.11.040

Hermes, M., & Stelling, D. (2016). Context matters, but how much? Latent state-trait analysis of cognitive ability assessments. *International Journal of Selection and Assessment, 24*(3), 285-295. https://doi.org/10.1111/ijsa.12147

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory and Cognition, 42*(3), 464-480. https://doi.org/10.3758/s13421-013-0364-z

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435-447. https://doi.org/10.3102/00028312021002435

Levacher, J., Koch, M., Hissbach, J., Spinath, F. M., & Becker, N. (2021). You can play the game without knowing the rules - but you're better off knowing them. *European Journal of Psychological Assessment 38*(1), 15-23. https://doi.org/10.1027/1015-5759/a000637

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*(6), 1672-1682. https://doi.org/10.1037/0021-9010.92.6.1672

Lievens, F., Sackett, P. R., & Zhang, C. (2021). Personnel selection: A longstanding story of impact at the individual, firm, and societal level. *European Journal of Work and Organizational Psychology, 30*(3), 444-455. https://doi.org/10.1080/1359432X.2020.1849386

Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence, 37*(4), 412-421. https://doi.org/10.1016/j.intell.2009.03.011

Matton, N., Vautier, S., & Raufaste, É. (2011). Test-specificity of the advantage of retaking cognitive ability tests. *International Journal of Selection and Assessment, 19*(1), 11-17. https://doi.org/https://doi.org/10.1111/j.1468-2389.2011.00530.x

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*(2), 191-216. https://doi.org/10.1037/0033-2909.89.2.191

Potočnik, K., Anderson, N. R., Born, M., Kleinmann, M., & Nikolaou, I. (2021). Paving the way for research in recruitment and selection: Recent developments, challenges and future opportunities. *European Journal of Work and Organizational Psychology, 30*(2), 159-174. https://doi.org/10.1080/1359432X.2021.1904898

Randall, J. G., & Villado, A. J. (2017). Take two: Sources and deterrents of score change in employment retesting. *Human Resource Management Review, 27*(3), 536-553. https://doi.org/10.1016/j.hrmr.2016.10.002

Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). Wiley.

Salgado, J. F. (2017). Using ability tests in selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention.* (pp. 115-150). Wiley Blackwell. https://doi.org/10.1002/9781118972472.ch7

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44-66. https://doi.org/10.1016/j.intell.2018.01.003

Traut, H. J., Guild, R. M., & Munakata, Y. (2021). Why does cognitive training yield inconsistent benefits? A meta-analysis of individual differences in baseline cognitive abilities and training outcomes. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.662139

Tu, Y. K., & Gilthorpe, M. S. (2007). Revisiting the relation between change and initial value: A review and evaluation. *Statistics in Medicine, 26*(2), 443-457. https://doi.org/10.1002/sim.2538

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352-402. https://doi.org/10.1037/a0028446

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 445-471. https://doi.org/10.1146/annurev-orgpsych-032516-113349

Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business and Psychology, 31*(2), 233-248. https://doi.org/10.1007/s10869-015-9408-7

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology, 29*(1), 64-77. https://doi.org/10.1080/1359432X.2019.1681401

Zinn, F., Goerke, P., & Marggraf-Micheel, C. (2020). Selecting for cockpit crew. In R. Bor, C. Eriksen, T. L. Hubbard, & R. King (Eds.), *Pilot selection: Psychological principles and practice* (pp. 21-34). CRC Press, Taylor & Francis Group.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*(4), 399-413. https://doi.org/10.1037/1082-989x.12.4.399